Machine Learning 101

Bill Macready

NASA/RIACS
wgm@email.arc.nasa.gov

What's Ahead?

- Operational definition of learning
  - Types of learning
- Some general issues:
  - Probabilistic inference as a framework
  - Overfitting
  - Ensemble models
- Specific examples:
  - Unsupervised learning
  - Supervised Learning
- References for starting points to learn more

What is Learning?

- Learning = improving with experience at some task
  - improve at the task T
  - according to some measure of performance P
  - based on experience E
  - Example 1 – playing checkers
    - T: playing checkers
    - P: % games won
    - E: past games (may be against self)
  - Example 2 – prediction
    - T: forecasting unknown outcomes
    - P: prediction error
    - E: historical record
  - Example 3 – understanding
    - T: make sense of some data
    - P: how well the model describes the data
    - E: historical data

Types of Machine Learning

- Reinforcement Learning
  - Maximize some future-discounted reward signal in an unknown and noisy environment
- Supervised learning
  - Given labeled data predict label of previously unseen data
- Unsupervised learning
  - Given some unlabeled data try to make sense of it either by assigning labels or by building a probabilistic model which may have generated the model
    - Dimension reduction

Supervised Learning Formulation

- Standard formulation of problem
  - Given a dataset $D = \{x_i, y_i\}_{i=1}^{d}$ consisting of example input/output pairs try to learn the mapping $x \xrightarrow{f} y$
  - Input space can be
    - discrete/continuous/mixed
    - 1 or more dimensional
  - Output space can be scalar or vector and
    - Continuous: regression
      - e.g. time series prediction of stock price
    - Discrete: classification
      - e.g. is stock a strong buy, buy, hold, or sell?

Unsupervised Learning Formulation

- Standard formulation of problem
  - Given a dataset $D = \{x_i\}_{i=1}^{d}$ consisting of sample values try to build a model that would likely generate the observed data $P(x)$
  - Most commonly used unsupervised learning technique is clustering
  - Input space can be
    - discrete/continuous/mixed
    - 1 or more dimensional
  - Of course supervised learning can be understood as a particular instance of unsupervised learning (and vice versa)

Probabilistic Inference

Ames Research Center

- Most modern approaches to learning can be understood within a framework of probabilistic inference
  - The natural generalization of Aristotelian logic which reduces to logic when hypothesis are true or false
- 2 simple rules:
  - Sum rule: $p(A|I) + p(\neg A|I) = 1$
  - Product Rule: $p(AB|I) = p(A|B,I)\,p(B|I) = p(B|A,I)\,p(A|I)$
- Bayes theorem (which is just the product rule) governs how hypotheses are modified with data
  - Probability of hypothesis given data proportional to likelihood x prior:
  $$p(h|d) = \frac{p(d|h)\,p(h)}{p(d)}$$
- Pick the most likely hypothesis: $h^* = \arg\max_h \; p(h|d)$

Probabilistic Inference(2)

- The probabilistic models may be either
  - Parametric: shape of probability density specified a priori by some parameters,
    - e.g. linear regression where we parameterize in terms of slope, intercept and noise level
  - Non-parametric: use histograms or samples from probability density
    - e.g. particle filters
- Some machine learning approaches which appear not to have anything to do with probability theory are best understood as particular limits of probabilistic case
  - e.g. principal components analysis

Probability + Graphs = Efficiency

- Important to understand probabilistic independencies between random variables
  - E.g. two variables might appear to be correlated with each other and appear as p(a,b), but might actually be independent given a common underlying hidden (or latent variable), p(a,b|l) = p(a|l)p(b|l)
  - Can use such statements of independence to infer causal relationships
- Observations like this can great speed up calculations involving probabilities
- Bayesian networks are models of probability densities with conditional dependencies annotated as a directed graph for efficient processing

Over-Fitting

- For both supervised and unsupervised learning an important issue to be aware of is over-fitting
  - Given 10 data points fitting a ninth degree polynomial will almost never result in good predictions for unseen points
- Can be understood in terms of a bias/variance tradeoff:
  - Bias: measures the quality of the match between the model and the underlying truth
  - Variance: measures the specificity of the match

$$E_D\left[\{g(x;D)-F(x)\}^2\right]=\left[E_D\{g(x;D)-F(x)\}\right]^2+E_D\left[\{g(x;D)-E_D[g(x;D)]\}^2\right]$$

- Many parameter models can drive down the bias but will usually increase the variance
- Prior beliefs on parameters or the number of parameters can be used to alleviate this problem

Cross Validation

- To ameliorate overfitting a common technique is called cross validation
- In cross validation you only use a sample of the full data set to build the model and you hold out the rest to test the error of your model on the held out set
- Common used to set the best parameters of learning algorithms

Ensemble Modeling

- If we have a number of models each making predictions how might we combine them to form a single best guess and how good can this guess be?
  - For regression might take the average guess, this is good because it drives down the variance while leaving the bias unaffected and thus results in lower error rate
  - For classification we can take a vote for each class and go with the winner
- Other more sophisticated techniques also empirically appear to work quite well and there are the beginnings of theoretical understanding;
  - E.g. boosting: build new models in regions where old models are performing poorly

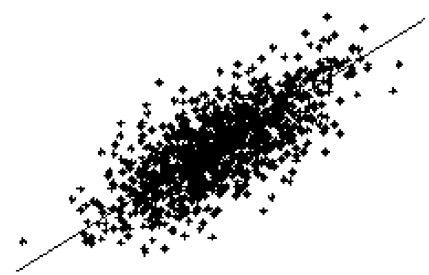
Some Example Methods:
Unsupervised

Methods of Unsupervised Learning

- Principal Component Analysis (PCA)
  - Assume the data have correlations and estimate the pairwise correlation matrix from the data:

$$\mu = \sum_{i=1}^{d} x_i, \quad C = \sum_{i=1}^{d} (x_i - \mu)^T (x_i - \mu)$$

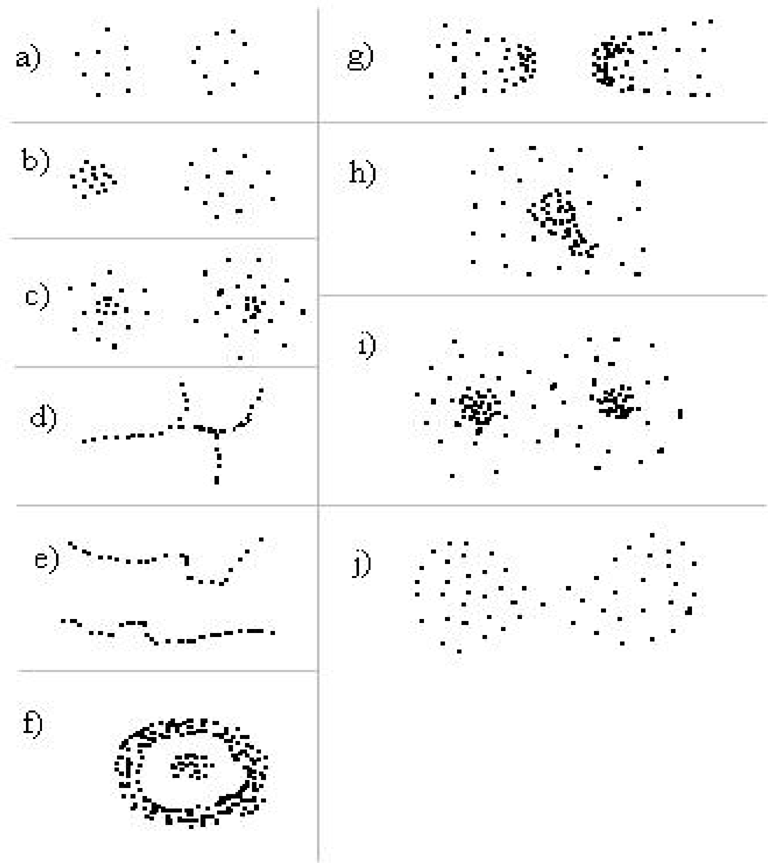
  - Diagonalize the covariance matrix to find directions (i.e. linear combinations of the variables) accounting for most of the variation
  - By disregarding the directions across which there is little variation we can reduce the dimensionality of the problem
  - Resulting variables will be uncorrelated



Data cloud with the first principal component drawn

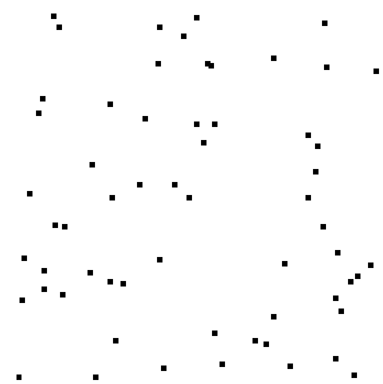
Methods of Unsupervised Learning

- Clustering or segmentation
  - People can do the job easily (at least in 2 dimensions) – but how??
  - Same concepts in higher dimensions

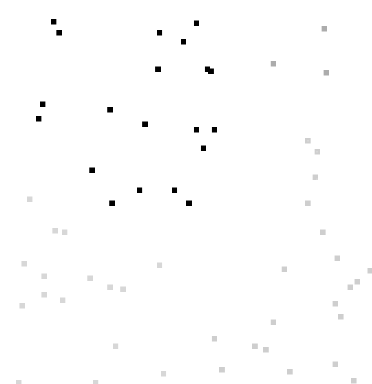


Methods of Unsupervised Learning

- K-means Clustering
  - Assumes there are k clusters and finds cluster centers such that maximum or average distance to the centroid of each cluster is minimized (mediod if data is binary)
  - Distance can be measured by a variety of means, e.g. Euclidean, correlation coefficient, Manhattan distance, etc



Original data



5 clusters (maximum distance)

Methods of Unsupervised Learning

- Gaussian mixture modeling
  - Assumes data generated by a mixture of Gaussian distribution:

$$p(x;\rho_1,\rho_2,\mu_1,\sigma_1,\mu_2,\sigma_2)=\rho_1 N(x;\mu_1,\sigma_1)+\rho_2 N(x;\mu_2,\sigma_2)$$

  - Finds all parameters by maximizing the likelihood of the observed data

$$L(\rho_1,\rho_2,\mu_1,\sigma_1,\mu_2,\sigma_2)=\prod_{i=1}^{d} p(x_i;\rho_1,\rho_2,\mu_1,\sigma_1,\mu_2,\sigma_2)$$

  - Membership to a cluster now becomes fuzzy, e.g. a point may be assigned 80% to cluster 1 and 20% to cluster 2
  - Likelihood can be optimized nicely with a procedure called EM or Expectation Maximization
    - Powerful technique with numerous applications
    - Converges to a local maximum of the likelihood function

Methods of Unsupervised Learning

- Other topics
  - Independent component analysis: separating conversations at a cocktail party
  - Topographic maps (Kohonen maps): finding two dimensional representations of higher dimensional data for visualization
  - Sparse bases: finding an overcomplete basis for the data so that any datum can accurately be represented with only a few basis vectors
  - Latent variable models: posit a few latent variables accounting for the data and infer these hidden variables

Some Example Methods:
Supervised

Methods of Supervised Learning

- Bayesian methods for inferring parameters in parametric models
  - Outlined earlier, simplest example is linear regression. The squared error is actually the negative log likelihood under a Gaussian approximation to the errors
  - Zillions of examples depending of the form of the model and the dependencies between variables.
    - more sophisticated but common example: hidden Markov models used in time series
    - Applies to both classification and regression
- Bayesian networks are a huge research topic now

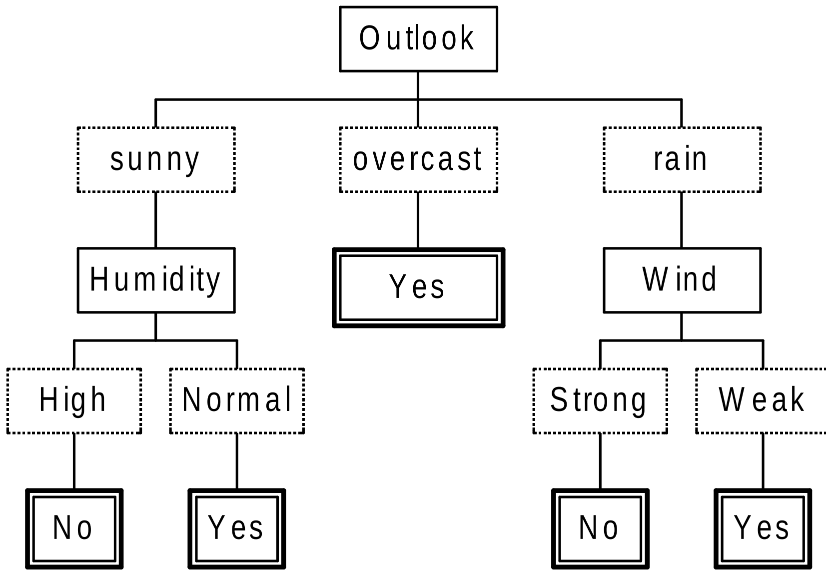
Decision Trees for classification

training data:

Day	Outlook	Temp	Humidity	Wind	Play Tennis
1	sun	hot	high	weak	no
2	sun	hot	high	strong	no
3	cloud	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	cloud	cool	normal	strong	yes
8	sun	mild	high	weak	no
9	sun	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sun	mild	normal	strong	yes
12	cloud	mild	high	strong	yes
13	cloud	hot	normal	weak	yes
14	rain	mild	high	strong	no

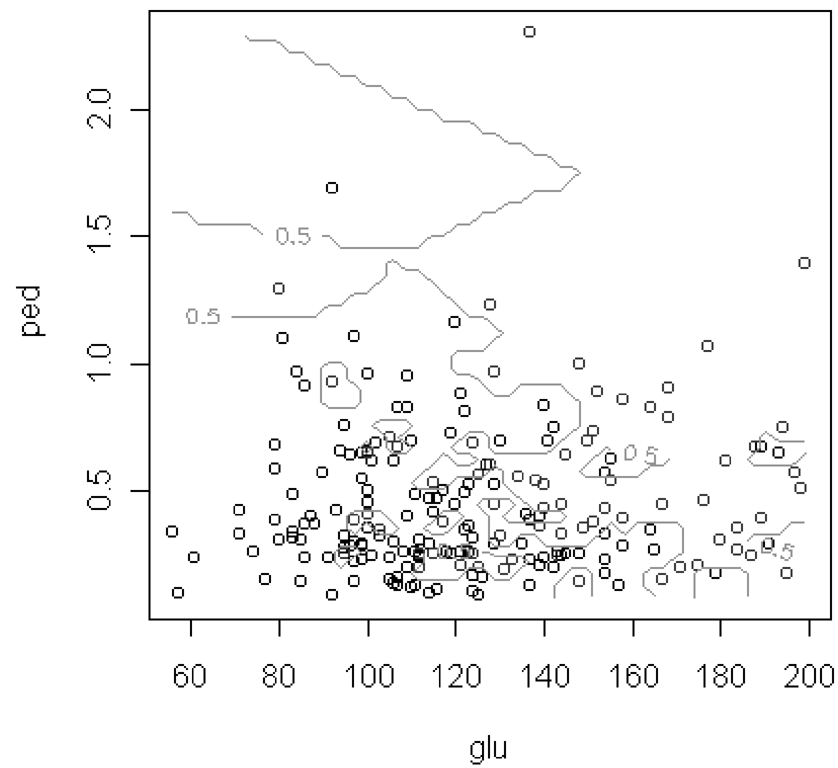
Methods of Supervised Learning

- The tennis decision tree
  - Determine those variables that reduce uncertainty the most and split on those
  - Trees can be pruned to limit overfitting
  - Very easy to interpret
  - Boundaries between classes are rectangular



Methods of Supervised Learning

- Nearest Neighbor rules
  - E.g. for classification find the "nearest" point in the training set of examples and say the new point is in the same class as that training example
  - Boundaries are more general and can be polygonal
  - Can use more than the single nearest point by selecting some number of nearby points and voting



Methods Of Supervised Learning

- Kernel-based Learning
  - If x is a new point at which we want to predict y then (x, y) should be "similar" to $(x_1, y_1), \ldots, (x_N, y_N)$
  - Similarity measures
    - For outputs: use a loss function, e.g. $L(y, y') = (y - y')^2$ for regression
    - For inputs: a kernel $k(x, x')$
  - The kernel is symmetric and generalizes the usual similarity measure on $\mathbb{R}^n$ which is the inner product $x \cdot x'$
    - In fact we can assume $k(x, x') = \Phi(x) \cdot \Phi(x')$, i.e. an inner product in some feature space H_k, therefore must be positive definite
    - Feature space often much higher dimensional

Methods of Supervised Learning

A little bit more formally

- The minimizer of

$$\frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i)]^2 + \lambda g_2(\|f\|_{H_k})$$

is of the form

$$f(x) = \sum_{i=1}^{N} c_i k(x_i, x)$$

where it is easy to find the expansion coefficients c_i by solving a quadratic programming problem

- Can generalize the loss term from quadratic to $g_1(y, f(x))$ and result still holds

Methods of Supervised Learning

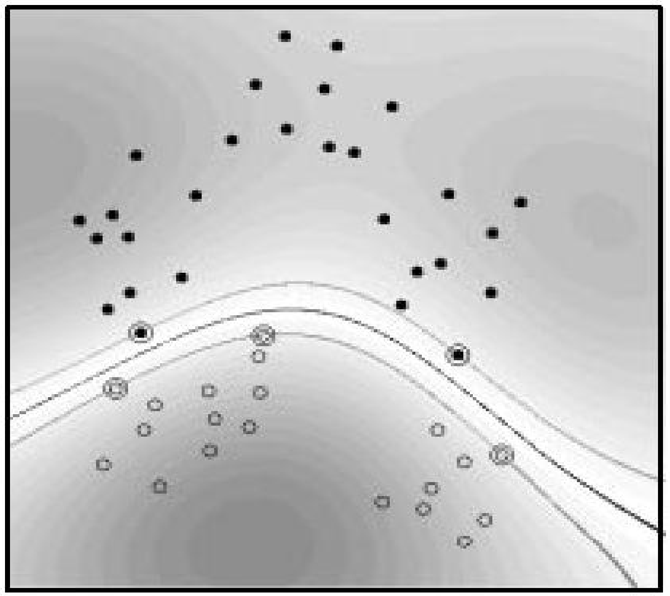


Figure 3: Example of a Support Vector classifier found by using
function kernel $k(x, x') = \exp(-\|x - x'\|^2)$. Both coordinate axes
to +1. Circles and disks are two classes of training examples; the
the decision surface; the outer lines precisely meet the constraint (2
the Support Vectors found by the algorithm (marked by extra ci
centers of clusters, but examples which are critical for the given
task. Grey values code the modulus of the argument $\sum_{i=1}^{m} y_i \alpha_i \cdot$
the decision function (34).)

Methods of Supervised Learning

- Many other methods
  - Neural networks: can approximate any function using an accumulation of many local non-linearities
    - Many variants of neural networks
  - Splines
  - Gaussian processes

Where to go for more information

- Overall survey introductions to machine learning:
  - *Pattern Classification*: Richard Duda, Peter Hart, David Stork, (Wiley 2000).
  - *Elements of Statistical Learning*: Trevor Hastie, Robert Tibshirani, Jerome Friedman, (Springer Verlag 2001)
  - NATO Advanced Study Institute in Learning Theory: http://www.esat.kuleuven.ac.be/sista/natoasi/ltp2002.html
- Probabilistic Inference:
  - *Probability Theory: the logic of science*, Edwin Jaynes, to be published April 2003
- Bayesian Networks
  - *Bayesian Networks and Decision Graphs*, F. V. Jensen (Springer. 2001)
- Kernel Methods:
  - An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Nello Christianini, John Shawe Taylor, (Cambridge University Press, 2000)
- Neural Networks:
  - *Neural Networks for Pattern Recognition*, Christopher Bishop, (Oxford University Press, 1995). Includes great accompanying Matlab software which is available online!